

Control of tongue movements in speech: the Equilibrium Point Hypothesis perspective

Pascal Perrier, H el ene L evenbruck and Yohan Payan

Institut de la Communication Parl ee, URA CNRS 368, Institut National Polytechnique de Grenoble & Universit e Stendhal, 46 Avenue F elix Viallet, 38031 Grenoble Cedex 1, France

Received 21st September 1994, and in revised form 1st July 1995

In this paper, the application of the Equilibrium Point Hypothesis—originally proposed by Feldman for the control of limb movements—to speech control is analysed. In the first part, physiological data published in the literature which argue in favour of such control for the tongue are presented and the possible role of this motor process in a global control model of the tongue is explicated. In the second part, using the example of the acoustic variability associated with vowel reduction, we focus on how the Equilibrium Point Hypothesis could help to search for physical regularities associated with a phonological sequence produced under variable speech conditions: the equilibrium point sequence could be invariant while the level of cocontraction and the timing of the commands could vary with the speech condition.

1. Introduction

A classical debate on coarticulation in speech is whether it is planned or due to dynamic properties of the articulatory apparatus. Articulatory data collected over the last 30 years support the idea that coarticulation is, at least in part, centrally planned—see Henke (1966), Bengu el & Cowan (1974), Perkell & Matthies (1992), Abry & Lallouache (in press) for anticipatory lip protrusion, Wood (1994) for tongue movement anticipation, or Hamlet & Stone (1981) for jaw movement anticipation. However, other works suggest that some coarticulation can be explained by the properties of the peripheral speech apparatus ( ohman, 1967; Fowler, 1977; Bell-Berti & Krakow, 1991). Therefore, it seems obvious that a speech production model should be able to separate central processes from dynamic effects.

In order to understand and to describe how central planning processes can operate, we adopt the approach suggested by Jordan (1990) (see also Jordan & Rumelhart, 1992). Based on learning processes and optimisation of kinematic criteria, Jordan's model is able indeed to supply successive intended vocal tract configurations for a given speech sequence that account for voluntary anticipatory and carry-over effects. The aim of this paper is to study how the dynamic properties of the articulators can influence the actual final articulatory configurations and make them different from the intended ones, in relation to prosodic effects. For that we

propose a model for speech motor control applied to a simple dynamic modelling of the articulators.

As for modelling of the tongue, complex biomechanical models were elaborated in the past years, based on precise descriptions of the anatomical and muscular structures (Henke, 1966; Perkell, 1974; Kiritani, Miyawaki & Fujimura, 1976; Kakita, Fujimura & Honda, 1985; Wilhelms-Tricarico, 1995). These models present a relatively high level of complexity, justified by the idea that only an accurate description of the real system can lead to a good understanding of speech production mechanisms. For example, Perkell justified his fundamental work on tongue modelling in this way: "It is intended to go beyond Henke's model principally by having an internal structure and function that is based to a greater degree on principles of motor control and neuromuscular function. Therefore it should provide the most natural possible framework for the exploration of physiological phenomena and testing of physiological and linguistic hypotheses." (Perkell, 1974). The control mechanisms of these accurate models were not the focus of the previous works and the inputs to these models usually consisted of measurable EMG signals.

With such an approach, Perkell proposed interesting hypotheses likely to explain the different tongue shapes associated with main vowels, and also to understand their configurational control (Perkell, 1990; this volume). However, as for inferences on the control of the spatio-temporal coordination of the articulators, no significant results were obtained. To our knowledge, with the exception of the appealing simulations of tongue movements obtained from EMG signals (Kakita *et al.*, 1985; Wilhelms-Tricarico, 1995), no work based on this modelling approach can presently describe the generation of speech movements. This situation explains why the main advances in the domain of coarticulation modelling were finally obtained with fairly simple mechanical descriptions of the articulatory system (e.g., Öhman, 1967; Kelso, Saltzman & Tuller, 1986; Saltzman & Munhall, 1989; Browman & Goldstein, 1990). In such works, dynamic aspects are essentially described using a second-order model, whose characteristics allow correct fitting of the kinematic data measured for human speech movements (Ostry, Keller & Parush, 1983; Nelson, 1983; Ostry & Munhall, 1985). Interesting results have been thus obtained, however, the global account of the dynamic behaviour of the articulators is not satisfactory. Indeed, the kinematic aspects due to dynamic properties of the speech apparatus cannot be discriminated from those of central control strategies specifically used for speech. The elaboration of accurate biomechanical models of the articulators remains, therefore, necessary, but with proper control variables.

EMG activations, and hence muscular force levels, seem not to be suitable control variables, since they are the consequences of an interaction between central and reflex activations to the motoneuron (MN) pool (Feldman, 1986). Moreover, the muscular structure of the speech apparatus is complex. It seems thus unrealistic to assume that each muscle may be centrally commanded individually, as would be the case in an EMG control model. In the case of the tongue, whose shape depends on approximately 20 muscles, the number of combinatory possibilities for muscular coordinations is quite high. It is, therefore, necessary to select control variables able to adequately represent the central commands controlling the synergies between muscles.

In this perspective, the Equilibrium Point Hypothesis (EP Hypothesis) proposed

Control of tongue movements in speech

by Feldman (1966; 1986) for the control of skeletal muscles is very appealing. In this paper, we will first present this theory and discuss whether the basic principles of the EP Hypothesis are compatible with the neurophysiological properties of the tongue and are able to give an account of muscle synergies. Then simulations using a simple mechanical model will be presented and discussed, emphasizing the benefit of the EP Hypothesis in understanding speech variability.

2. EP Hypothesis and tongue control

The basic assumption of the EP Hypothesis is that movement arises through changes in neural control variables that shift the equilibrium point of the motor system. It is based on elementary physiological mechanisms: motor innervation to skeletal muscles arises from α MNs which innervate the main body of the muscle, and from γ MNs which contribute to α MN excitation through reflexes; Feldman's suggestion is that the control variables are independent changes in the membrane potentials of α and γ MNs, which establish a threshold muscle length (λ) where the recruitment of MNs begins. If the actual muscle length is larger than λ , muscle activation, and hence force, vary in relation to the difference between actual and threshold muscle lengths, following a non-linear relationship (Feldman & Orlovsky, 1972). Each value of λ is related to a unique force-length relationship. The central specification λ can therefore be interpreted as the selection of a specific force-length relationship for the muscle. For a multimuscle system, the choice of the λ s determines in a unique way the spatial position of the mechanical equilibrium.

Moreover, a given equilibrium position can be specified by different combinations of λ s. These λ combinations build up a subset in the λ space which is characteristic of this equilibrium position and for which the synergies between muscles are implicitly taken into account (Feldman, Adamovich, Ostry & Flanagan, 1990). Modifying the values of λ s within a subset induces no movement, but changes the force distribution among the muscles and alters the global force level (cocontraction level). Shifting the equilibrium point, in the space of the degrees of freedom, corresponds to moving, in the space of the central commands, from one λ subset to another. Hence, in this theory, muscles are not controlled individually. Rather central control variables are specified with regard to the kinematic degrees of freedom.

Consequently, movement control consists of specifying, in the space of the degrees of freedom, a virtual trajectory in terms of successive equilibrium positions. From simulations performed with a two-joint arm model based on the EP Hypothesis, Flanagan, Ostry & Feldman (1993) suggested that complex arm trajectories, measured for human reaching movements to fixed and suddenly displaced visual targets, can be obtained with simple virtual trajectories, corresponding to constant rate equilibrium shifts towards a final equilibrium position. In a target related conception of speech control (see below), this proposal of a simple virtual trajectory linking successive equilibrium positions, in the space of the degrees of freedom, is particularly appealing: target undershoot phenomena are indeed currently observed in speech movements, and such a proposal offers a way to understand how the measured articulatory trajectories can be related to the intended targets underlying the movement. However, the notion of simple virtual trajectory is a matter of controversy in the literature. Katayama & Kawato (1993) for instance,

by using parallel inverse statics and inverse dynamics models, suggest that the linear trajectories observed in the space of the degrees of freedom for fast or low-stiffness point-to-point movements are accounted for with very complicated virtual trajectories. Obviously such simulations are very dependent on the characteristics of the dynamic modelling, and especially on the muscle models, which are explicitly different in the two considered works. In the absence of any decisive evidence, we will propose, following Flanagan *et al.* (1993), a model for speech control based on simple virtual trajectories, linking successive targets.

In summary, the EP Hypothesis presents interesting features likely to contribute to the elaboration of an efficient control system for speech articulators: it is physiologically founded; the neural control variables are related to physical characteristics (Equilibrium Point) in the space of the degrees of freedom of the system; each equilibrium position has a specific projection in the space of the control variables (λ subsets); synergies between muscles are implicitly taken into account. A jaw/hyoid bone model was thus proposed by Laboissière, Ostry & Feldman (in press) shedding light on the relations between the motor control space and the degrees of freedom space. Before proposing a model of the tongue based on the EP Hypothesis, the question of whether neurophysiological processes involved in tongue movements present compatible characteristics must be addressed. A short description of the innervation of the tongue seems thus necessary.

2.1. Neurophysiology of the tongue

Most of the oral mucosa (and hence the tongue surface) is rich in many different types of mechanoreceptors. These receptors respond to various kinds of mechanical distortion arising, for instance, from contact between the tongue and the palate or teeth. Their response consists of generating a depolarising current in the sensory fibre. They are not evenly distributed throughout the oral region. Grossman (1964) describes a progressive decrease in the density of sensory endings from the front to the rear of the mouth. This progression is particularly noticeable in the tongue, where the tip seems better endowed with sensory receptors than any other part of the oral system.

Beside these receptors, which are situated in the mucosa throughout the oral region, there are muscle spindles within the tongue musculature. These receptors provide information on length and rate of length change in the muscle, and therefore act as essential elements in a servo-mechanism system by means of the stretch reflex loop. Cooper (1953) also suggests that there is a non-uniform distribution of muscle spindles in the tongue. She found most spindles in the superior longitudinal muscle near the midline and in the front third of the tongue, and in the transverse muscle in the mid-region towards the lateral borders. Walker & Rajagopal (1959) also found neuromuscular spindles in the genioglossus, hyoglossus, styloglossus, and in the intrinsic muscles of three newborn infants. They observed that the genioglossus contains the greatest number of spindles. A relatively greater density of muscle spindles has thus often been found in parts of the muscles that are thought to require fine adjustments in the production of complex articulations (such as [s], [f], [i] or [e]). This supports the idea that muscle spindles could play an important role in the control of tongue movements.

Adatia & Gehring (1971) have investigated the proprioceptive sense of the

Control of tongue movements in speech

tongue in twelve human subjects. The tongue was held by an operator and moved in various directions at random. Subjects were asked to determine the direction of these passive movements. Eleven of the twelve subjects had no difficulty, while only one subject said he could not “feel where the tongue was” when it was moved upwards. This work shows the presence of a proprioceptive feedback in the control of tongue position, in accordance with previous experiments done by Weddel, Harpman, Lambley & Young (1940). The afferent source for this feedback could be the muscle spindles as proposed by Pearson (1945), Bowman & Combs (1969) and Fitzgerald & Sachithanandan (1979), and the afferent information could, according to the same authors, be sent back through the hypoglossal nerve, a nerve of the lingual musculature which is often still described as purely motor.

Even if the presence of stretch reflexes in the human tongue musculature has not yet been demonstrated (Neilson, Andrew, Guitar & Quinn, 1979), neurophysiological data on the tongue lead us to think that proprioceptive afferent information could be sent by the muscle spindles to the spinal bulb. This scheme would then be likely to provide the afferent facilitation mentioned in Feldman’s theory (1966).

2.2. Which role for the EP Hypothesis in a general control model of the tongue?

The central problem in developing a speech production model is the question of its ability to explain first, how articulators are recruited for the production of a given linguistic unit and second, how acoustic and articulatory features associated with the same linguistic unit can vary according to phonological and phonetic contexts.

It is well known that several articulatory positions can produce the same sound (Atal, Chang, Mathews & Tukey, 1978; Gay, Lindblom & Lubker, 1981; Maeda, 1990; Boë, Perrier & Bailly, 1992). This compensation ability can be exploited by speakers to deal with imposed or chosen constraints such as a pipe in the mouth or a will to produce clearly visible articulatory movements in a noisy speech condition. This freedom leads to different speaker-dependent strategies in normal speech such as speaker-dependent anticipatory phenomena (Abry & Lallouache, in press).

Clearly, the EP Hypothesis is not aimed at the simulation of such phenomena. For that, it may be proposed, following Jordan (Jordan, 1990; Jordan & Rumelhart, 1992), that a speaker uses an internalised representation (forward model) of the relations between articulatory positions and the relevant acoustic features. This representation gives an account of all possible articulatory positions associated with the same sound. In this way, a correspondence can be found between a sound sequence and a kind of temporal multidimensional ribbon in the articulatory space. The choice of an articulatory path within this temporal ribbon could result, following Lindblom (1988, 1990), Keating (1988) or Jordan (1990), from a balance between speaker-oriented principles, such as, for example, the minimisation of a global potential energy along the articulatory path, and perceptive requirements (Laboissière, Schwartz & Bailly, 1991). This would imply the definition, at the level of the Central Nervous System (CNS), of intended articulatory trajectories for which, depending on the context, different articulatory positions could be associated with the same linguistic unit (planned coarticulation). The EP Hypothesis could be involved at this stage of speech production control.

The EP Hypothesis is suited indeed to describe how the intended trajectory will actually be achieved by the articulatory apparatus with its own inertial properties

and force generation principles, and how motor commands can be related to the phonological level. By defining movement as a result of shifts from posture to posture, the EP Hypothesis allows us to generate a continuous articulatory trajectory from a succession of discrete commands which can be related to a similarly discrete phonological sequence. Accordingly, articulators move towards targets—defined in terms of equilibrium positions in the space of the degrees of freedom—that are predetermined by the planned coarticulation, with dynamic properties that are dependent on biomechanical and force generation features.

A comparable description of the target as an attractor in the articulatory space is proposed by researchers at Haskins Laboratories (Saltzman, 1986; Kelso, Saltzman & Tuller, 1986). A number of differences between the Haskins model and the EP Hypothesis should be emphasized. In the Haskins model, the attractor acts directly in the geometrical task space defined by the vocal tract variables (lip, glottis, and velum apertures, tongue body and tongue dorsum constrictions). The problems with this suggestion are the following:

(1) There is no underlying mechanism for control.

(2) The position of each articulator is inferred from the trajectory towards an attractor in the task space following coordinate structures which account for the relations between articulatory positions and vocal tract variables. These relations are purely geometrical; they correspond to the geometrical articulatory model of the vocal tract initially elaborated by Mermelstein (1973) and developed by Rubin, Baer & Mermelstein (1981).

Articulatory trajectories are, therefore, dependent on properties of the dynamic attractors in the task space and on the respective weights of the different vocal tract variables (cf. Saltzman & Munhall, 1989). No account is given of either the inertial or the muscle mechanical properties of each articulator.

(3) The objective of the task is defined in the geometrical space with no consideration whatsoever of the perceptive space.

In contrast, in our modelling the EP Hypothesis enables the control of an articulatory model in which each articulator, characterised by its muscle's mechanical and inertial properties, moves towards a target, which is defined in relation to the perceptive space (planned coarticulation) and which is interpretable in terms of neural commands.

3. EP Hypothesis and speech variability

Besides the features already mentioned in the introduction in favour of a control model of tongue movements based on the EP Hypothesis, Feldman's hypothesis (Feldman, 1966) might shed light on an issue in speech variability, namely *prosodic* variability such as stress and speaking rate. By assuming that both the equilibrium position of the articulator and the muscle cocontraction level are controlled by neural commands, the EP Hypothesis clearly differentiates the objective to be reached by the articulators (the target defined by the equilibrium point) from the way to reach this objective (the force level and the timing of equilibrium shifts). Our aim is thus to propose a model for the control of tongue movements that describes how articulatory objectives are encoded and how the movement towards such objectives is parameterised.

Control of tongue movements in speech

3.1. A simple tongue model based on the EP Hypothesis

As mentioned in the introduction, present tongue models are very sophisticated, and to build a comprehensive model of their control in speech represents a long and complex task. We propose to work with a much simpler mechanical model, and to study how a control based on the EP Hypothesis can help in the analysis of speech variability.

Maeda's approach (Maeda, Honda & Kusawaka, 1993), which groups muscles by their common specific action on the tongue, seems quite adapted to our task. These authors claim indeed that "although the tongue muscular system is anatomically complex, it is organised into a small number of functional blocks for speech production." They showed that the tongue position for a vowel can be determined by two sets of antagonistically paired EMG activities. The hypoglossus (HG) and the genioglossus posterior (GGp) function symmetrically: the activation of GGp corresponds to a displacement of the tongue forwards and upwards, while the HG activation induces a backward and downward displacement. Similarly, the styloglossus (SG) and the genioglossus anterior (GGa) correspond to antagonistic actions: backwards and upwards for the former and forwards and downwards for the latter. Following these results, we consider that shape and position of the tongue body are influenced by two independent lingual articulators and by the jaw. Each of these articulators is supposed to be controlled by a pair of antagonist muscle sets.

Following classical modelling (e.g., Cooke, 1980), we use a second-order system to model the actions of each pair of opposing muscles. An articulatory degree of freedom is thus represented by two springs connected together, one representing the group of agonist muscles and the other the antagonist ones (Perrier, Abry & Keller, 1989). For each degree of freedom (i), the second-order model of tongue articulators is described by the following equation, normalised by the mass:

$$(1) \quad \ddot{y}_i + f_i \dot{y}_i + K_i (y_i - y_{ei}) = 0$$

K_i is the sum of the stiffnesses, normalised by the mass, of the two springs and corresponds to a global muscle cocontraction. K_i is assumed to remain constant during each speech sequence. We chose a damping value f_i characteristic of a slight undercritical damped system, i.e., below $2\sqrt{K_i}$, so as to allow target positions to be reached fast enough. We thus set f_i to an *ad hoc* value: $1.89\sqrt{K_i}$. Note, however, that the sensitivity of the system to this parameter is small within the range $]1.6\sqrt{K_i}, 2\sqrt{K_i}[$. Variable y_{ei} specifies the spatial equilibrium position of the system, and can thus be considered as the spatial target towards which the movement is intended. Our basic idea for speech control is that, for each degree of freedom, a specific equilibrium position is associated by the CNS with each phoneme within a sequence, thus defining the successive targets of the movement. The equilibrium shift from one target to the next is not abrupt, and the transition times between the targets can also be modified by the controller. The central commands consist then of the specification of successive equilibrium points, of the level of cocontraction, and of the successive transition- and hold-times of the temporal equilibrium trajectories. Fig. 1 gives an example of the evolution of the equilibrium position for the production of three phonemes.

In summary, in our perspective, the targets of the movement are specified by y_e , while the dynamical characteristics of the movement are parameterised by the

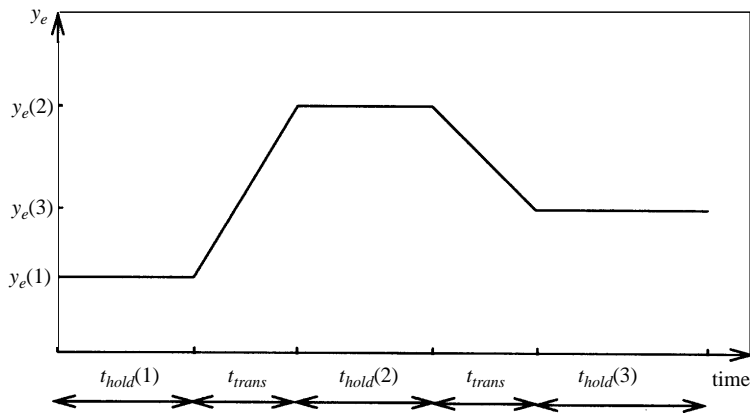


Figure 1. Trajectory of the equilibrium position y_e for a sequence of three phonemes.

cocontraction level (K) and the timing of the central commands. In this framework, our hypothesis is that speech variability associated with prosodic effects can be simulated for an equilibrium point sequence by adjusting only the cocontraction level and/or the timing of the central commands.

3.2. An example of speech variability: vowel reduction

Lindblom (1963) observed variations in formant frequencies associated with speech rate modifications as well as with different degrees of stress for vowel /u/ in three different consonantal contexts ([dud], [bub] and [gug]). Represented in the vowel triangle F_1 - F_2 , both prosodic changes correspond to a shift of the formant pattern from one edge of the triangle (standard /u/ position) towards the central part (vowel reduction).

Different models have been proposed during the past few decades, arguing the nature of the original cause of this phenomenon. Lindblom's first suggestion rests on the hypothesis that speech production consists of achieving successive targets related to successive phonemes. In this perspective, vowel reduction would correspond to an undershoot of the intended vowel target, for which Lindblom proposed a quantitative prediction from 3 factors: adjacent consonantal context, intended vowel target, and vowel duration. According to this model, for a given context, duration reduction (either due to speech rate increase or to stress reduction) systematically prevents the articulatory system from performing the full, required gesture: both the articulatory and formantic trajectories undershoot their intended targets.

A number of studies emerged later revising this first assumption. Gay (1978) suggests that the "degree of [vowel] reduction is linked to stress, regardless of the relative or absolute duration of the segment." This was confirmed for vowels /i/, /a/, and /u/ by Engstrand (1988) who found that spectral characteristics were significantly influenced by stress but not by speech rate. In the same vein Nord (1986) noted that unstressed vowels coarticulate strongly with their context, whatever the duration. It was, therefore, clear that the first duration-dependent undershoot model could not adequately describe empirical data. In 1992, Lindblom, Brownlee, Davis & Moon revised the original undershoot model by introducing

Control of tongue movements in speech

speech style (citation form, clear speech, etc.) in those terms: “reduction processes can be seen as contextual assimilations durationally induced, but, within certain limits, speakers appear capable of controlling the precise degree of reduction.”

This proposal is questioned by Van Bergem (1993) and Pols & Van Son (1993) who refute the hypothesis of a target-oriented speech production. Instead, they propose that speech production would consist of generating relevant features in the dynamic part of the formant trajectories. According to them, vowel duration should thus not be considered the original cause for vowel reduction, but as one of the consequences of the transition control associated with speech style: “The stressed vowel tokens were generally longer and less reduced [...] than the unstressed ones [...]. However vowel duration alone was not enough to explain those differences. It is probably the other way round: stress, context and speaking style result in certain formant and duration changes, and are for the greater part actively controlled by the speaker.” (Pols & Van Son, 1993).

In this open debate, our aim is to propose a quantitative modelling of target-oriented speech production and to assess to which extent speech variability can be generated from invariant targets by controlling duration, context, and speech style.

For that, an acoustic corpus was recorded which presents clear vowel reduction phenomena, in vowel transitions essentially involving tongue front/back movements. Associated articulatory trajectories were then inferred from the acoustics by inverting an articulatory model of the vocal tract, which describes the relations between the articulatory and acoustic spaces in speech. The inputs to the tongue model described above were then optimised in order to correctly fit tongue movement. Finally, acoustic variability was generated by acting on cocontraction level and on timing of the commands.

3.3. Methodology

3.3.1. The corpus

The corpus consists of the sequence [iai] in the French carrier sentence “il y a immédiatement” recorded for a native French male speaker (Beautemps, 1993). Three different speech conditions are studied involving variations of speech rate and stress. It should be noted that “stress” here means *focus* put on a specific vowel. Under the first condition—slow and stressed—the speaker was asked to speak slowly and to stress the vowel [a]; under the second condition—slow and unstressed—the instructions were to speak slowly with no specific stress on [a]; finally, under the third condition—fast and stressed—the instructions were to speak at a fast rate with stress on [a].

Fig. 2 shows the formant patterns extracted from the signal, under the three speech conditions. The narrow space between formants F_1 and F_2 is characteristic of vowel [a]. This space tends to increase under the second and third conditions: a vowel reduction phenomenon is thus observed.

3.3.2. Recovering of central commands from the acoustic signal

Recovering of central commands from the acoustic signal involves, in our approach, two successive inversion procedures. The first one could be described as a kinematic

